리즘이다. 웹 문서의 하이퍼링크는 자신과 주제가 같은 상호 관련성을 가지는 다른 문서와 링크로 연결된다. 따라서 하이퍼링크를 이용한 웹 문서 분류 기법은 URL의 hostname을 추출하여 분류한다.

```
Feature = Documen_Feature_Extraction();
Documen_Length = GetDocumen_Length();
Documen_Classification = TRUE;

if(FEATURE_MINIMUM_NUMBER <= (Feature/ Documen_Length))
        Document_Resemblance_Measurement(Feature);
else
{
    while(GetDocument_Hyperlinks_Url()>0)
    {
        Feature = Feature + Hyperlink_Document_Feature_Extraction();
        Documen_Length = Documen_Length + GetHyperlink_Documen_Length();
    }
    if(FEATURE_MINIMUM_NUMBER <= (Feature/ Documen_Length))
        Document_Resemblance_Measurement(Feature);
    else
        Documen_Classification = FAILURE;
}
```

그림 2. 효용성 및 문서유사도 측정 알고리즘

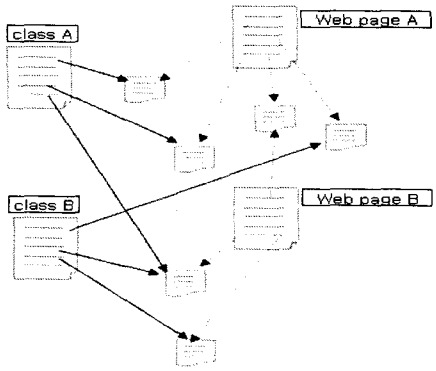

그림 3. 하이퍼링크를 이용한 문서 분류

문서의 유사도 측정은 문서 분류에 가장 많이 쓰이는 Naive Bayesian[7]을 이용한다.

그림 3은 하이퍼링크를 이용한 웹 문서 분류 방식으로 Class A, Class B는 학습에 의해 생성된 것으로 각각 3개의 링크를 가지며, 웹 문서 분류 시 기준에 된다. Web page A는 분류를 위한 새로운 질의 문서로 링크 파싱 결과 4개의 링크를 가진다. 각각의 링크들 중 두 개는 Class A에 포함되고, 한 개는 Class B에 포함되므로, Web page A는 Class A로 분류 된다. Web page B의 경우는 Class A와 1개의 링크가 연결되며, Class B와 2개의 링크가 연결되므로 Class B로 분류된다.

표 1은 하이브리드 방식의 웹 문서 분류기의 문서 분류 단계를 나타낸 것으로 학습 단계와 분류 단계로 나누어진다.

본 논문에서 제안하는 하이브리드 방식의 문서 분류기는 모두 학습 과정을 가진다. 텍스트 정보를 이용한 텍스트 기반의 웹 문서 분류기의 학습과정은 주제와 관련된 문서에서 특징을 추출하여 Class를 생성하며, 생성된 Class를 이용하여 문서를 분류하게 된다.

분류기 학습
1. 주제중심의 웹문서 분류기의 학습과정
1.1 사용자가 제공하는 문서를 이용하여 학습
1.1.1 기존의 검색 엔진을 통하여 주제어를 검색
1.1.2 검색 결과에서 주제 관련 문서들에 대한 특징 추출
2. 하이퍼링크를 이용한 분류기의 학습과정
2.1 사용자가 제공하는 URL을 이용하여 학습
2.1.1 사용자가 제시하는 URL에서 URL 추출
2.1.2 추출된 URL에서 새로운 특징을 가지는 URL추출
2.1.3 새로운 특징을 가지는 URL이 있을 경우 2.1.1부터 다시 시작
웹 문서 수집
3. 웹 문서 수집을 시작할 URL 입력
4. 입력받은 문서의 하이퍼링크 URL 추출
5. 추출된 URL을 방문할 목록을 저장하는 큐에 입력
6. 큐에서 방문할 URL 획득
7. 획득한 URL을 이용하여 웹 문서 수집
8. 문서분류
8.1 텍스트 데이터 추출
8.1.1 웹 문서의 텍스트 데이터 추출
8.1.2 수집된 텍스트 데이터가 충분할 경우 8.1.5로 이동
8.1.3 수집된 텍스트가 부족할 경우 하이퍼링크로 연결된 문서에서 텍스트 데이터를 보충
8.1.4 보충된 텍스트 데이터 역시 부족할 경우 8.2로 이동
8.1.5 주제중심의 웹 문서 분류기를 이용하여 문서 분류 후에 9번으로 이동
8.2 하이퍼링크를 이용한 분류기로 분류
8.2.1 웹 문서의 하이퍼링크 URL을 이용하여 문서분류
8.2.2 분류를 실패할 경우 하이퍼링크로 연결된 문서안의 URL로 다시분류시도
8.2.3 하이퍼링크로 연결된 문서안의 URL로 문서 분류 못할 경우 분류 실패로 판정
9. 분류가 끝날 경우 분류한 웹문서의 URL을 추출 후에 5번으로 이동
10. 하이퍼링크를 이용하여 분류에서 분류가 실패 했을 경우 6번으로 이동
11. 큐에 방문할 URL이 없을 경우 프로그램 종료

표 1. 하이브리드 방식의 웹 문서 분류기의 구동 방식

또한 하이퍼링크를 이용한 분류기는 주제와 관련된 URL을 이용하여 연관된 URL을 수집하고 각각의 Class를 생성한다.

4. 실험

본 논문에서는 다양한 형식의 웹 문서를 분류 하기위해 다음과 같이 정의한다.

웹 문서가 텍스트와 이미지로 혼합된 경우 분류 주제는 "웹 크롤러 관련 서적" 과 "뉴스정보" 로 하고, 이미지나 멀티미디어가 텍스트 보다 많은 경우 분류 주제를 "여

행" 으로 한다. 텍스트가 많은 문서의 경우 "웹 크롤러" 라는 주제를 이용하였다. 각 주제는 3가지 형식의 웹 문서를 대표하는 것으로 "웹 크롤러 관련 서적" 과 "뉴스정보" 는 이미지 데이터와 텍스트 데이터가 혼재되어있는 형식의 웹 문서이다. 이것은 이미지 데이터와 텍스트 데이터가 혼재되어있는 웹 문서 형식에서 어떠한 분류 방식이 좋은 효과를 보이는지를 실험하는 환경이며, "여행" 에 대한 주제를 사용할 경우 사진 이미지 또는 여행 장소에 대한 이미지 데이터가 많으므로 이미지 데이터가 70%이상인 웹 문서 형식에 적합하다. "웹 크롤러" 라는 단순 주제의 경우 수식 설명 또는 논문 등과 같은 텍스트 데이터가 70% 이상인 텍스트 환경이므로 텍스트 문서 형식에 대한 실험 환경에 적합하다. 주제 분류를 위한 시작 페이지는 야후에서 주제어를 이용하여 검색된 문서 중 주제 유사도가 가장 높은 문서로 설정한다. 실험 후 수집된 문서가 주제와 부합되는 웹 문서인지를 판단하는 기준은 주제에 대하여 기존에 알려진 웹 문서를 이용하여 수집된 문서가 주제와 부합되는지 판단한다. 주제와 부합되는지 판단이 되지 않는 웹 문서에 대해서는 직접 문서를 확인하는 과정을 통하여 주제와 일치하는지 판단한다.

본 논문에서는 다양한 형식의 웹 문서 환경에서 문서 분류를 위해 기존의 텍스트 정보를 이용한 경우, 하이퍼링크를 이용한 경우, 그리고 제안된 하이브리드 방식의 경우를 비교 실험 하였다.

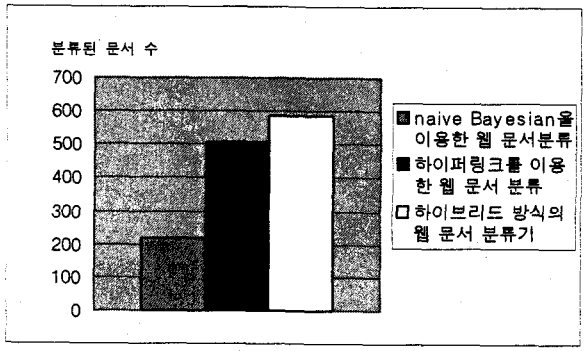

그림 4. 이미지 문서에 대한 분류결과

그림 4는 이미지 데이터가 70% 이상인 웹 문서들을 분류한 결과이다. 같은 시간동안 텍스트 정보를 이용한 텍스트 기반의 웹 문서 분류기는 217개의 문서를 수집 하였으며, 하이퍼링크를 이용한 분류기는 505개의 문서를 수집하였다. 그리고 하이브리드방식의 분류기는 587개의 문서를 분류하였다. 문서 형식에서 텍스트 데이터 보다 이미지 데이터가 많은 경우 텍스트 데이터를 이용한 텍스트 기반 주제 중심 분류방식이 하이퍼링크와 하이브리드 방식보다 분류률이 떨어진다. 하이퍼링크를 이용한 웹 문서 분류기는 URL의 주제를 이용하므로 텍스트를 이용한 분류기보다 288개의 문서를 더 수집할 수 있었다. 하이브리드 방식의 분류기는 URL 분류 후 분류가 실패로 판정될 경우 다시 텍스트를 이용한 텍스트 기반의 주제 중심 문서 분류가 가능하므로 하이퍼링크를 이용한 경우 보다 82개의 문서를 더 많이 분류할 수 있었다.

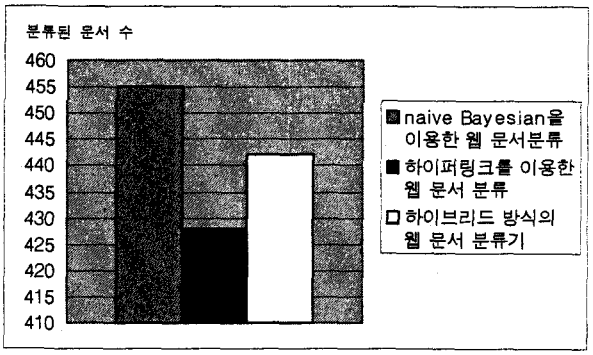

그림 5. 텍스트 문서에 대한 분류결과

그림 5는 텍스트 데이터가 70%이상인 웹 문서를 대상으로 실험한 분류 결과이다. 텍스트를 이용한 웹 문서 분류기는455개의 문서를 수집 하였고, 하이퍼링크를 이용한 분류기는 428개 문서를 수집하였다. 그리고 하이브리드방식의 분류기는 442개의 문서를 분류하였다. 텍스트 정보가 주를 이루는 문서의 경우 하이퍼링크를 이용한 경우보다 더 많은 문서를 수집하였다.

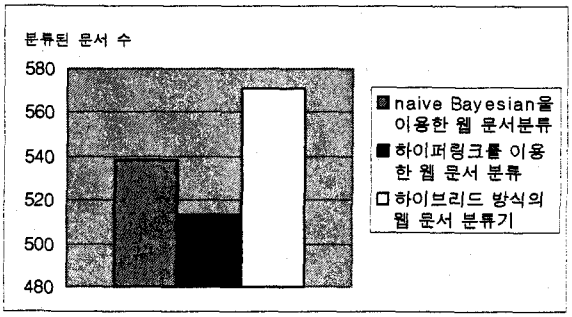

그림 6. 혼용 문서에 대한 분류결과

그림 6은 이미지와 텍스트 데이터의 비율이 일정하게 혼합되어 있는 웹 문서를 분류한 결과이다. 텍스트 정보를 이용한 웹 문서 분류기는 538개의 문서를 수집 하였으며 하이퍼링크를 이용한 분류기는 513개의 문서를 수집하였다. 그리고 하이브리드방식의 분류기는 571개의 문서를 분류하였다. 다양한 형식으로 구성된 문서를 분류하는 경우에는 하이브리드 방식의 이용한 분류기가 더 많은 문서를 수집 하였다.

그림 7은 웹 문서의 혼합된 형식의 문서를 분류할 경우 시간당 발생하는 오분류율을 측정한 것이다. Y축은 오분류로 판정된 문서의 수이며, X축은 측정된 시간이다.

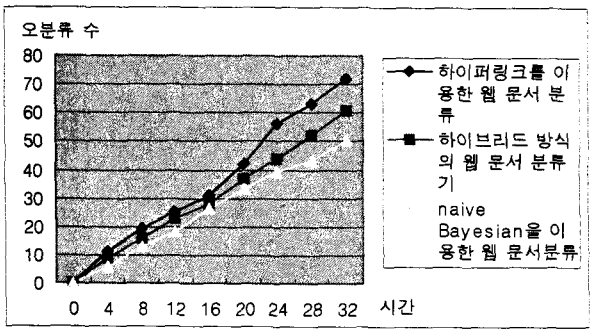

그림 7. 각 분류기의 오분류 측정

텍스트 정보를 이용한 웹 문서 분류기는 텍스트 형식의 웹 문서를 분류할 경우 분류율이 높다. 그러나 그림 4와 같이 이미지 데이터가 많은 경우 분류 능력이 현저히 떨어지는 것을 볼 수 있다. 또한 하이퍼링크를 이용한 분류에서는 텍스트 형식의 경우 18%에 가까운 오분류가 발생한다. 반면 하이브리드는 모든 형식의 웹 문서에서 고른 분류 결과를 보여주며, 낮은 오분류율을 나타낸다.

6. 결 론

인터넷의 발전과 다양한 웹 문서의 증가로 텍스트나 이미지, 멀티미디어 등을 이용한 정보 표현 형식으로 변화하고 있다. 기존의 웹 문서 분류기는 텍스트 형식의 웹 문서 분류를 중심으로 연구되었다. 그러나 기존의 방식은 텍스트 데이터를 이용하여 정보를 추출하고 분류하기 때문에 이미지 등을 이용하여 정보를 표현하는 환경에는 적합하지 않다. 또한 텍스트 데이터가 부족한 경우 오분류가 자주 발생하며, 분류 능력이 떨어지는 문제점을 발생한다.

따라서 본 논문에서는 이러한 문제점을 해결하기 위해 텍스트 데이터와 하이퍼링크를 결합한 하이브리드 방식을 제안하였다. 하이브리드 방식은 텍스트 정보가 부족한 경우 하이퍼링크를 이용하여 재분류 할 수 있기 때문에 오분류율이 감소하며, 웹 문서에서 이미지나 텍스트가 차지하는 비율과 상관없이 모든 형식의 웹 문서를 분류하는 것이 가능하다. 결과적으로 기존의 분류 방식이 갖는 문제점을 해결하고 분류기의 성능을 개선할 수 있었다. 향후 다양한 실험을 통하여 분류방식에 맞는 평가함수를 찾는다면 더욱 향상된 분류기를 개발할 수 있을 것이다.

참고문헌

[1] A. McCallum, "Building Domain-Specific Search Engines with Machine Learning Techniques," Proceeding AAAI Symp. Intelligent Agents in Cyberspace, AAAI Press, pp.28-39, 1999.
[2] J. Cho, "Efficient Crawling through URL ordering," Computer Networks and ISDN Systems, Vol.30, pp.161-172, 1998.
[3] M, Hersovici, "The SharkSearch Algorithm-An Application: Tailored Web Site Mapping," Proceeding of 8th Int' l World Wide Web conference, pp.213-225, 1998.
[4] Paul De Bra, "Information Retrieval in Distributed Hypertexts," Proceeding of 4th RIAO Conference, 1994.
[5] S. Chakrabarti, m. Ven den Berg And B.E. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," WWW-8. 1999.
[6] S. Chakrabarti, m. Ven den Berg And B.E. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," WWW-8. 1999.
[7] 조창희, 주변정보 분할을 이용한 주제 중심 웹 문서 수집기, 한국정보과학회, 제12-B권 제6호 통권 제102호, pp. 697-702, 2005.
[8] 박민규, 유태명, 김준태, 웹사이트의 분류와 필터링을 위한 하이퍼링크 정보의 효용성에 관한 연구, HCI'99 학술대회 학술발표논문집, pp.94-98, 1999.
[9] http://www.mochanni.com/
[10] http://www.webcrawler.com/